

Клименко С.В, Рыков В.В.

Институт физики высоких энергий

Институт языкознания РАН

klimenko@sirius.ihep.su

rykov@iling.msk.su

Метафора моря и первый закон Муэрс

Стало уже общим местом говорить, что Интернет предоставил море информации почти каждому жителю Земли. И огромная часть этих жителей, впервые получая из этого моря свой стакан морской воды, вполне этим довольна. И это действительно так. Китайская философия говорит, что истина выглядит парадоксально. И в то же время эта парадоксальность является движущей силой прогресса в соответствующей области.

Дело в том, что после первой эйфории часто происходит отрезвление и множится число недовольных. Начнем с простейших информационных потребностей. Действительно, если пользователь Интернета запросил телефон жителя Нью-Йорка, то ему Интернет выдаст адреса 25 бесплатных телефонных книг. Не все от этого приходят в восторг. Номер телефона они не узнали.

Возьмем более сложный случай. Один из авторов наблюдал следующий факт. В Сети нужно было найти важную информацию. И было известно, что она там есть. На ее поиск два квалифицированных специалиста затратили два рабочих дня. И это тоже не послужило поводом для восторга.

Действительно, с появлением Интернета разверзлись многие информационные шлюзы между пользователями и информацией, стало казаться, что "смазались" многие информационные закономерности. Но это не так. Произошел тектонический сдвиг, но законы остались те же. Пользователю, выразившему свою информационную потребность в английской фразе, дадут море информации. Далее - смотри выше.

Здесь мы подходим к первому закону Муэрс. Интернет его не отменил и он гласит: "Потребитель уклоняется от использования информационно-поисковой системы (ИПС)

всякий раз, когда для него более затруднительно и хлопотно искать информацию, чем обойтись без нее" (Mooers, 1960). Другими словами, коммуникация с ИПС (в нашем случае с Интернет) прерывается.

Семиотика ИПС и второй закон Муэrsa

Можно сделать вывод, что как массовые пользователи ИПС (имеется в виду прежде всего Интернет), так и профессиональные сообщества специалистов не вполне удовлетворены существующим положением вещей и активно требуют его улучшения. Одним из ответов на эту потребность стали так называемые информационные аукционы (Кавасаки, 2000). Клиенты присылают в специализированную фирму в Сети свои запросы и сумму, которую они готовы заплатить за нужную информацию (хотя она, возможно, и так доступна). И это нормально. Бесконечное разнообразие рынка информационных потребностей порождает соответствующую гамму ее поставщиков.

Для точного, научно обоснованного поиска нужной информации необходимо строить специализированную семиотическую (знаковую) систему. Ее элементами будут: информационно-поисковый язык (ИПЯ), на котором составляется так называемый поисковый образ документа (ПОД). ПОД является сложным знаком, соотношенным с содержанием документа. Также на ИПЯ формируется другой сложный знак - поисковый образ любого пользовательского запроса (ПОЗ). Извечная проблема ИПС, сформулированная выше, звучит как проблема релевантности ПОД ПОЗ. Интернет не смог этого отменить. Просто массовому пользователю предоставлена возможность сформулировать свой информационный запрос на естественном языке (ЕЯ). И уже интернетовская "искалка" генерирует свой ПОЗ на основе полученной от пользователя информации. Что из этого зачастую выходит, очень кратко сказано выше.

Что делать для повышения эффективности ИПС? Второй закон Муэrsa гласит, что надо переходить от слов к понятиям. Это очень напоминает современную уголовную доктрину жизни "по понятиям", но только на уровне лексики, что может послужить первой иллюстрацией того, о чем пойдет речь. Хорошей второй иллюстрацией будет поиск в Интернете с запросом в виде заголовка этой статьи. Кроме информации по корпусной лингвистике будут выданы документы о религиозной организации Corpus Christi и о корпусе российских войск в Туркестане. Не надо вникать в организацию конкретной интернетовской ИПС, выдавшей эту галиматью, чтобы сделать вывод, что строится она по словам, а не по понятиям.

Конечно, при известной, по выражению В.В.Виноградова, "филологической сноровке" эти трудности можно обходить. Многочисленные советы по этому поводу печатаются в компьютерной прессе. Муэрс же учит, что начинать надо с того, что каждое понятие надо описать особой лексической единицей - дескриптором (Mooers, 1972). Второй его закон звучит так: "Как цветовые таблицы служат средством обнаружения дальтонизма у человека, так и метод дескрипторов есть средство для выявления лиц, "привязанных к словам" (word-bound) или "невосприимчивым к идеям" (idea-blind). Выявление таких лиц имеет большое значение, потому что "привязанный к словам" человек, по-видимому, не способен обеспечивать информационный поиск, основанный на использовании идей (т.е. не зависящий от слов), а именно такой поиск и является наиболее свойственным и наиболее желательным для людей, которые в своем

мышлении не привязаны к словам (Mooers, 1960).

Однако Интернетом пользуются все - и дальтоники и люди "невосприимчивые к идеям". Более того - проблема состоит даже не в этом. Проблема состоит в трудоемкости семиотически правильного индексирования. То есть индексирования каждого документа, выставленного в Интернете, в терминах дескрипторного ИПЯ. Имеется в виду прежде всего автоматическое индексирование. Здесь просто неизбежен какой-либо компромисс. Одним из таких компромиссных подходов к решению проблемы является методология корпусной лингвистики.

Корпусная лингвистика и построение

семиотически адекватного ИПЯ

Одним из принципов корпусной лингвистики является антропоцентрический подход. Внимание фокусируется не на том, какими должны быть язык или речь (в данном случае - какой должен быть ИПЯ), а что собой реально представляет языковая личность, использующая ИПС, - такая, какая она есть (Рыков, 1999). В нашем случае это означает, что приходится считаться с тем, что наиболее массовым пользователем Интернет является неподготовленный пользователь.

Тогда задача, стоящая перед лингвистом, исповедующим принципы корпусной лингвистики, выглядит так. Следует построить ПОЗ, основанный на понятиях и выраженный на дескрипторном ИПЯ, имея в качестве исходного речевого материала запрос, сформулированный пользователем на лексическом уровне.

Для решения этой задачи на помощь приходит другой принцип корпусной лингвистики - опора на эталонный корпус текстов, адекватно отражающий данный тип коммуникации. Такой корпус помогает решить лингвистические трудности, исходя не из идеализированных представлений и моделей, а реального речевого материала, уже употреблявшегося в данном типе коммуникации.

Этот подход для решения задачи перехода от лексического способа формулировки запроса к понятийному поиску был реализован и описан уже несколько лет назад (Futrelle, Gauch, 1993). Реализовывается этот подход у авторов по следующему алгоритму. Сам процесс они называют расширением запроса (expanding query). Запрос расширяется и становится "концептуальным" (conceptual), используя обращения к оперативному корпусу необходимой лексики (online database) и так называемым специализированным "матрицам подобия" (similarity matrix). Таким способом устраняется бессмысленное лексическое сравнение (word matching) при поиске и осуществляется переход к концептуальному поиску (conceptual retrieval).

Здесь же авторы делают важное замечание. Как уже говорилось, для автоматизированного перехода от слов к понятиям служат матрицы подобия. Они расширяют исходную лексику пользовательского запроса до понятийного. В принципе, если такой переход лексически однороден, то достаточно одной матрицы подобия. Приходится же выбирать одну из нескольких. Другими словами, одно и то же слово будет по-разному расширяться и формироваться понятийный ПОЗ. В зависимости от того, в какой понятийной или специализированной области происходит поиск.

На понятийную область может указать снова эталонный корпус текстов, отражающий лексику соответствующей понятийной области. Если обратиться к приведенному выше примеру о лексическом поиске по заголовку этой статьи, то специализированная матрица

подобия должна сформировать такой ПОЗ, релевантными которому были бы только ПОДы документов с лингвистической информацией (по корпусной лингвистике).

Мы приходим к выводу, что автоматический переход в ИПС от лексической формы запроса к понятийному (дескрипторному) ПОЗ возможен. Он требует в качестве предварительной основы создания корпуса текстов, содержащего реальный речевой материал, использовавшимся в такого рода информационных запросах. То есть в речевой модели канала коммуникации, основанного на реально происходивших актах коммуникации. Алгоритмическое и программное моделирование такого подхода, который декларируется как корпусная лингвистика, показывает, что такой канал принципиально неоднороден. Это явление обусловлено гетерогенностью коммуникативной среды Интернет (Клименко, Крохин, 1997).

Выходом из этого противоречия может быть более строгое определение понятийных границ области поиска. То есть разбиение коммуникативной среды на однородные (гомогенные) среды с последующим моделированием каждой из них в отдельности.

Литература

Кавасаки Г. Выложиться в Web // Computerworld Россия. - М., 2000. - N.1. - С. 14-15.

Клименко С.В., Крохин И.В. и др. Электронные документы корпоративных сетях. - М., 1999. - 272 с.

Клименко С.В., Рыков В.В. Диалоговое извлечение знаний из корпуса текстов // Диалог-99. - Москва-Таруса, 1999.

Маурер Г. Упраление знаниями на основе Web-технологий // Computer Weekly. - 1988. - N.34-36. - С. 34-37.

Рыков В.В. Прагматически ориентированный корпус текстов // Диалог-99. - Москва-Таруса, 1999.

Рубашкин В.Ш. Представление и анализ смысла в интеллектуальных информационных системах. - М.:Наука, 1989.

Черный А.И. Заметки об информатике и дескрипторах // Азгальдов Э.Г. и др. Дескрипторный словарь по информатике. - М.: ВИНТИ. - 1991. - С.3-13.

Futrelle R.P., Gauch S. Experiments in syntactic and semantic classification and disambiguation using bootstrapping // Acquisition of Lexical Knowledge from Text.- Columbus, OH. Assoc. Computational Linguistics, 1993. - P. 117-127.

Mooers C.N. "Mooers" law, or why some retrieval systems are used and other are not // American Documentation. - 1960. - Vol.11, N.3.

Mooers C.N. Descriptors // Encyclopedia of library and information science / A.Kent and H.Lancour, eds. - Vol.7. - New York, 1972.- Vol.7. - P. 31-45.